IEEE SA
STANDARDS ASSOCIATION

INDUSTRY CONNECTIONS REPORT

IEEE SA INDUSTRY CONNECTIONS RESEARCH GROUP ON ISSUES OF AUTONOMY AND AI IN DEFENSE SYSTEMS

A FRAMEWORK FOR HUMAN DECISION-MAKING THROUGH THE LIFECYCLE OF AUTONOMOUS AND INTELLIGENT SYSTEMS IN DEFENSE APPLICATIONS

IEEE

How to cite this document: IEEE SA Research Group on Issues of Autonomy and AI in Defense Systems. (2024). *A Framework for Human Decision Making Through the Lifecycle of Autonomous and Intelligent Systems in Defense Applications*. New York, NY: IEEE SA.

AUTHORS

This report has been prepared by the following individuals (listed alphabetically), all participating in their individual capacities:

Sten Allik
Rachel Azafrani
David Barnes
Ingvild Bode
Ariel Conn
Eileen M. Lach
Craig Lennon
Rain Liivoja
John (Jack) Shanahan
Melodena Stephens
Lisa Titus
Alan R. Wagner

ACKNOWLEDGMENTS

Special thanks are given to the following current and former members and affiliates of the Industry Connections Research Group, all participating in their individual capacities, who attended and contributed to meetings and helped review this paper in its various iterations (in alphabetical order):

Greg Adamson, Chad Bieber, Justin Bronk, Simon Chesterman, Di Cooke, Christian Ellis, Noah Ford, Charles Ovink, Jonathon Parry, Giacomo Persi Paoli, Signe Redfield, Brendan Schulman, Yvette Stevens, Catherine Tessier, and Wendell Wallach.

On behalf of the Research Group, we extend our gratitude to IEEE SA and the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. We are particularly thankful to Konstantinos Karachalios for his vision, to Anja Kaspersen for her guidance, and to John C. Havens for his support. We also thank Andrew Wright for his editorial assistance. We would also like to express our sincere thanks to the dedicated reviewers who took the time and lent us their expertise to offer feedback on the various iterations of the document. On behalf of the Industry Connections Research Group, special thanks are also given to Ariel Conn for her significant contributions and commitment to the development process of this document. The Chairs for this process were Ingvild Bode and Rachel Azafrani. For questions about the report, please reach out to: iaaids@ieee.org.

This document was prepared using an open, collaborative, and consensus-building approach, following the processes of the Industry Connections program, a program of the IEEE SA.

TERMS OF USE

Users may access, download, print, and/or retain one (1) copy of the document for individual use only, including job-related functions.

Compliance with this framework does not constitute compliance with applicable legal and regulatory requirements.

Except as allowed by the copyright laws of the United States of America or applicable international treaties, or as explicitly allowed in these Terms of Use, copies of the document may not be further copied, prepared, and/or distributed without prior written permission from IEEE.

Report for training, teaching, or any further distribution or work product of the document, in whole or in part, must be licensed by IEEE, and a licensing fee will be applied.

Please contact us by completing the IEEE Standard Permissions Use Form.

The Institute of Electrical and Electronics Engineers, Inc. 3 Park Avenue, New York, NY 10016-5997, USA

PDF: STDVA27266 979-8-8557-1141-7

Find IEEE standards and standards-related product listings at: http://standards.ieee.org.

NOTICE AND DISCLAIMER OF LIABILITY CONCERNING THE USE OF IEEE SA INDUSTRY CONNECTIONS DOCUMENTS

This IEEE Standards Association ("IEEE SA") Industry Connections publication ("Work") is not a consensus standard document. Specifically, this document is NOT AN IEEE STANDARD. Information contained in this Work has been created by, or obtained from, sources believed to be reliable, and reviewed by members of the IEEE SA Industry Connections activity that produced this Work. IEEE and the IEEE SA Industry Connections activity members expressly disclaim all warranties (express, implied, and statutory) related to this Work, including, but not limited to, the warranties of: merchantability; fitness for a particular purpose; non-infringement; quality, accuracy, effectiveness, currency, or completeness of the Work or content within the Work. In addition, IEEE and the IEEE SA Industry Connections activity members disclaim any and all conditions relating to: results; and workmanlike effort. This IEEE SA Industry Connections document is supplied "AS IS" and "WITH ALL FAULTS."

Although the IEEE SA Industry Connections activity members who have created this Work believe that the information and guidance given in this Work serve as an enhancement to users, all persons must rely upon their own skill and judgment when making use of it. IN NO EVENT SHALL IEEE OR IEEE SA INDUSTRY CONNECTIONS ACTIVITY MEMBERS BE LIABLE FOR ANY ERRORS OR OMISSIONS OR DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO: PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS WORK, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE AND REGARDLESS OF WHETHER SUCH DAMAGE WAS FORESEEABLE.

Further, information contained in this Work may be protected by intellectual property rights held by third parties or organizations, and the use of this information may require the user to negotiate with any such rights holders in order to legally acquire the rights to do so, and such rights holders may refuse to grant such rights. Attention is also called to the possibility that implementation of any or all of this Work may require the use of subject matter covered by patent rights. By publication of this Work, no position is taken by the IEEE with respect to the existence or validity of any patent rights in connection therewith. The IEEE is not responsible for identifying patent rights for which a license may be required, or for conducting inquiries into the legal validity or scope of patent claims. Users are expressly advised that determination of the validity of any patent rights, and the risk of infringement of such rights, is entirely their own responsibility. No commitment to grant licenses under patent rights on a reasonable or non-discriminatory basis has been sought or received from any rights holder. The policies and procedures under which this document was created can be viewed at http://standards.ieee.org/about/sasb/iccom/.

This Work is published with the understanding that IEEE and the ICCom members are supplying information through this Work, not attempting to render engineering or other professional services. If such services are required, the assistance of an appropriate professional should be sought. IEEE is not responsible for the statements and opinions advanced in this Work.

TABLE OF CONTENTS

A FRAMEWORK FOR HUMAN DECISION MAKING THROUGH THE LIFECYCLE OF AUTONOMOUS AND INTELLIGENT SYSTEMS IN DEFENSE APPLICATIONS

EXECUTIVE SUMMARY

The framework set out in this document addresses stakeholders involved in policy, design, testing, procurement, decision-making, deployment, and evaluation processes related to autonomous and intelligent systems (AIS), especially in public-sector decisions about defense applications.

This document highlights that there is no such thing as a completely autonomous system. Humans are always involved somewhere along the way. The framework, therefore, aims to inform discussions about manifold technical, ethical, and legal issues arising out of pursuing a human-centric approach across the entire lifecycle of AIS, especially with regard to human responsibility and accountability.

The scope of this report is broad, encompassing diverse insights that reflect on technical, ethical, and legal considerations inherent in AIS with a focus on military settings. The framework supports stakeholders in raising and offering first steps towards applying existing sets of broad ethical principles and standards in the context of AIS, including but not limited to those associated with Article 36 of Additional Protocol I to the Geneva Conventions.

Considering the full lifecycle of systems—from development to procurement, testing, deployment, use, sustainment, and decommissioning—the framework offers a granular view of stakeholder involvement, clarifying when and how challenges should be addressed and identifying who should be responsible and accountable at each lifecycle stage.

Decision-makers should consider that AIS may not necessarily be appropriate for use across defense settings, whether because the system itself is immature or the level of human knowledge is insufficient. They need to comprehensively assess whether the system is ethical and legal. This may include considering issues, such as the potential benefits, risks, and harms of developing and deploying any system; identifying potential blind spots; effects on and gaps in human skills; technical or scientific shortfalls; and issues, such as supply chain reliability.

The framework also calls attention to how many risks are introduced and must be addressed at the early stages of the lifecycle, including at the design stage of AIS.

The framework is the outcome of discussions among the IEEE SA Industry Connections Research Group on Issues of Autonomy and AI in Defense Systems, an independent group with members drawn from various countries, academic disciplines, and professional backgrounds. It can serve as a first step towards the potential development of standards processes around AIS in defense with a focus on ethically aligned design.

Nine lifecycle stages and five ongoing activities

Members of the Research Group concluded that there are nine stages in the lifecycle of AIS in defense contexts (illustrated in FIGURE 1), in which human decision-makers play especially key roles, and between which there is likely to be a transfer of responsibility and accountability to new human decision-makers:

a) Before AIS development
   - General legal, ethical, and related technical concerns identified and addressed
   - Rationale for military development and use of AIS, formulation of system requirements, and considering the role of researchers

b) Research and development
c) Procurement and acquisition
d) Testing, evaluation, verification, and validation (TEVV)
e) Considering the human: Education, training, and human-system integration
f) Political and strategic considerations
g) Operational-level command and control
h) Tactical employment
i) Review, reuse, and/or retire

Additionally, the group identified five “ongoing activities” in which human decision-making elements occur repeatedly throughout the lifecycle:

1) Evaluation of legal, ethical, and policy concerns
2) Responsibility, accountability, and knowledge transfers
3) Considering the human: training, education, and human-system integration
4) TEVV, monitoring, hardware system or software updates and interoperability, maintenance
5) Risk assessments

Some of the nine lifecycle stages are necessarily sequential—for example, a system must be developed before it can be deployed. Many, however—especially during the first half of the lifecycle—may have significant overlap, and may start or end at around the same time. These stages may also be repeated many times after a system has been deployed, when it received new hardware or software updates. Redundancy is built into the framework by design.

The focus throughout is on the humans involved in decisions. The framework is designed to help identify the relevant decision-makers, their role at each stage, the knowledge and documentation they need from previous decision-makers, and the types of questions and issues to be addressed at each stage.

Everyone associated with deploying an AIS for military purposes needs to have a common understanding of its capabilities and limitations. Individual, incremental steps in design and development may seem unproblematic, but then trigger cascading effects when combined with other technological components. A system cannot be understood in isolation from other technical components, the logistical environment of use, the human environment, or the personnel it will interact with for a given task.

Lessons from the process

The Research Group approached the task of developing the lifecycle framework by first creating a fictional example scenario involving an uncrewed underwater vehicle deployed to protect ships carrying food in a conflict situation. They defined possible outcomes of this deployment and worked backward to explore, test, and validate stages in the lifecycle as they developed it.

This approach proved valuable in helping to differentiate stages that were more chronological from stages that needed to be addressed throughout most of the lifecycle. It clarified the need to address questions at appropriate stages, the trade-off between apparent simplicity of operation at later stages and complexity of development at earlier stages, and the importance of providing adequate technical and legal documentation at each stage for the use of people involved at later stages.

The process of developing the framework made clear the need to involve technical people from civilian sectors in discussions with experts in legal and ethical issues, policymakers, and military leadership. Different experts have different understandings of terminology and capabilities. Different stages of the lifecycle tend to have a preponderance of one type of expert rather than the diversity of experts needed for a more sufficient awareness of the multiple concerns AIS raises.

Given how contentious discussions about AIS in military contexts can be, there was an encouraging degree of agreement among Research Group members on many issues. Areas of broad agreement included the value of considering a risk-based framework—accounting for differences in factors such as predictability, type of target, duration, and environment of use—and the importance of commanders and operators having sufficient contextual understanding of how a system works and what it can and cannot do.

A contribution to an ongoing debate

Debates about the legality—and, to a more limited extent, the ethics—of AIS in military settings have been taking place for around a decade. Some uses are relatively uncontroversial—such as navigation, management, analytics, and logistical assistance—although the systems' accuracy, interoperability, ability to mitigate against algorithmic bias, and verifiable maturity can still be questioned. Other capabilities remain highly contentious, especially those that would allow a system, as defined by the International Committee of the Red Cross (ICRC), to determine its own objectives: identifying, selecting, and applying force to targets without human intervention.

Nothing in this framework should be taken as precluding that new regulations, standards, policies, or laws may need to be developed to address concerns about AIS in defense contexts. The framework is intended to advance and contribute to these discussions and efforts. Although this framework was developed for AIS in defense, it is broad and adaptable enough that it should prove useful for stakeholders in the development and use of most AIS.

1. BACKGROUND

1.1. ABOUT THIS PROJECT

The framework set out in this document is intended to help inform a range of stakeholders involved in various types of decision-making related to autonomous and intelligent systems (AIS), especially in the military context. The framework provides a granular means for identifying the technical considerations that are necessary to meet legal requirements, to inform and guide ethical discussions, and to determine human responsibility and accountability across the entire lifecycle of AIS.

Article 36 of Additional Protocol I to the Geneva Conventions requires states to review the legality, under international law, of all new weapons, means, and methods of warfare—which increasingly incorporate AIS. In practice, given the new and technical complexity of these systems, states often do not have the requisite technical knowledge and procedures in place to conduct meaningful, standardized reviews. The framework set out in this document is intended, in part, to contribute to filling that gap. However, its scope is much broader.

Many AIS used in defense contexts are not codified as weapons, means, or methods of warfare, and in many cases, may not be different from commercial versions of AIS. Yet states will nonetheless need to address various legal, ethical, and technical challenges of their use. It is, for example, possible for systems to be compliant with international law but still harmful. Existing sets of principles on ethics often use terms that can be difficult to operationalize, such as systems being “responsible” or “trustworthy.” The way such terms are commonly understood in relation to human processes does not always map easily onto computational systems, especially but not only in military contexts.

The framework is intended to offer more technical guidance around putting principles into practice. While focused on military contexts, the framework can also be used beyond. It helps identify and address issues related to AIS, notwithstanding whether the systems are for military or civilian use. Therefore, it is intended for all stakeholders and decision-makers across the lifecycle stages, including those in research and development and procurement, policymakers, and leaders engaged in decisions about the use of force and the use of AI in public sector operations, and participants in international governance and treaty discussions.

The framework is the outcome of discussions among the IEEE SA Industry Connections Research Group on Issues of Autonomy and AI in Defense Systems (“the Research Group”). With members drawn from a range of nations and disciplines, the Research Group’s mandate included defining the next steps for addressing the challenges

identified in a 2021 document produced by a predecessor independent research group under IEEE SA auspices, entitled *Ethical and technical challenges in the development, use, and governance of autonomous weapons systems*.[1]

The Research Group agreed on the approach of developing a framework that could more clearly define the points of human decision-making throughout the lifecycle of AIS in defense contexts. The Research Group provided a technical platform for diverse stakeholders to share insights and develop common understanding.

In developing the framework, the Research Group had in mind systems that are in development now or may be developed in the next few years. The framework is intended for all public sector systems that incorporate autonomous or AI capabilities. For example, the development of self-driving, self-maneuvering, or self-navigation capabilities would benefit from adherence to this framework. The Research Group also designed the framework so that it could adapt as technology changes in the future.

1.1.1. A NOTE ON TERMINOLOGY

The title of the Research Group—"On Issues of Autonomy and AI in Defense Systems"—reflects its wide scope. Discussions about such issues can often be complicated by the use of terminology that is defined differently by different organizations and can be interpreted too narrowly.[2]

This document eschews, where possible, the use of "autonomous weapons systems (AWS)" to reduce the risk of excluding relevant defense and military systems. The preferred terminology is AIS in defense or military contexts. This has the advantage that AIS—autonomous and intelligent systems—is now an industry conformity discourse, and this document aims to encourage discussion of conformity as it applies to public sector settings.

The Research Group also considered alternatives to the word "lifecycle," which some consider inappropriate for systems designed to injure and kill. As "lifecycle" suggests a biological entity that can be born, live and die, it may contribute to the problem of anthropomorphizing technology and exaggerating the extent to which AI systems process information in an analogous way to humans. Suggested alternatives included "idea to

[1] After several years of detailed expert consultations, IEEE SA was invited to provide insights from industry practices, ontological standards for autonomous and intelligent systems, processes, relevant research, and strategies for incorporating safety and ethics by design. This was initiated by the Director of Disarmament in Geneva and the Chair of the Group of Governmental Experts for Emerging Technologies in the area of Lethal Autonomous Weapons Systems.

[2] The Research Group considered putting forward its own definition of AWS. As many other groups have already offered varying definitions, it is not clear there is value in adding yet another definition to the discussion. In international treaty discourse, for instance, the acronym "LAWS"—lethal autonomous weapon system—is often used. However, this can exclude AIS systems not considered to be "lethal," which is a matter of interpretation. Other processes address this concern by referring instead to "AWS," or autonomous weapon systems, but this may also be overly restrictive. Some interpret "autonomous" as implying a total lack of human input, rather than the presence of some degree of autonomous or AI-enabled capabilities. Similarly, the term "weapons" can be narrowly read as excluding other relevant military or defense systems.

decommission stages" and "conception to decommission stages." Ultimately, however, the Group decided to stick with "lifecycle" as it is so commonly used in the technology and military fields.

In many of the lifecycle stages described below, the use of some specialist technological and military language is unavoidable. In general, however, the document aims to use terminology that is as accessible as possible to stakeholders who may not have a background in either the technology or defense sectors, such as politicians, policymakers, ethicists, and lawyers.

1.2. A BRIEF OVERVIEW OF THE DEBATE AROUND AIS IN DEFENSE SETTINGS

Nothing in this framework should be taken as precluding that new regulations, policies, or laws may need to be developed to address concerns about AIS in defense contexts. The framework is intended to advance and contribute to ongoing debates on this subject while addressing the reality of AIS developed for military purposes. Even for relatively uncontroversial uses of AIS, stakeholders should follow this framework so that development and use address legal and ethical concerns, e.g., that such systems are not misused later in ways that could have been anticipated. The Research Group also recognizes that new issues will arise as technology changes.

AIS can offer the same kind of attractions to political and military leaders as such systems do to industry leaders. For example, AIS promises to enable more efficient operations, improve cost-effectiveness, take over routine work, and free up skilled humans for other tasks. Systems are increasingly becoming smaller, more affordable, and more expendable. However, the hidden costs associated with developing AIS are frequently overlooked, including, for example, environmental costs. AIS is developed to become more capable of undertaking a wider range of tasks in various domains of warfare: air, space, ground, sea, and non-kinetic, such as information and communications.

Some uses of AIS in defense settings are relatively uncontroversial—such as navigation, management, analytics, and logistical assistance—although the systems' accuracy, interoperability, ability to mitigate against algorithmic bias, and verifiable maturity can still be questioned. Other capabilities remain highly contentious, especially those that would allow a system, as defined by the International Committee of the Red Cross (ICRC), to determine its own objectives: identifying, selecting, and applying force to targets without human intervention.

Some military personnel and political leaders argue that such capabilities could, in theory, lead to fewer civilian casualties by lowering the possibility of error in high-stress environments. However, as AIS can fail in different

ways than humans, it cannot be assumed that a reduction in errors would necessarily imply that a system is safer. Concerns include:

- Unanticipated failure modes are possible as systems move from lab testing to use in real environments. Systems may react unexpectedly to natural variations in data, such as blur in images; have unforeseen vulnerabilities to attacks against data, networks, or models; or be susceptible to "reward hacking," which results in subtle changes to what a system is trying to achieve.
- Systems may prove less able than humans to anticipate potential breaches of International Humanitarian Law (IHL) or the Law of Armed Conflict (LOAC). Some AI systems have been shown to exacerbate existing human biases and discrimination. It can be difficult to understand how AI arrived at a decision (the "black box" problem). Humans tend to place too much trust in machines ("automation bias").
- Loss of human control over the use of force raises ethical concerns related to human dignity and moral agency. When a system behaves unexpectedly, neither those who developed its component parts nor those who decided to deploy it may feel morally responsible. It may be difficult to hold specific humans accountable for actions performed by machines.

Debates about the legality—and, to a more limited extent, the ethics—of AIS in military settings have been taking place for around a decade across national and international bodies, and within governments and militaries. These debates are becoming more urgent. As states anticipate the growing use of such systems by opposing forces—including non-state actors and individuals, as the technology becomes more accessible—there is potential for arms-race dynamics and unregulated proliferation.

Systems are at significant risk of being designed, developed, and deployed without sufficient human and regulatory oversight. One reason is that the secrecy and sensitivity of such projects limit willingness to seek feedback. Another is that individual, incremental steps in design and development may seem unproblematic, but then trigger cascading effects when combined with other technological components.

TABLE 1 summarizes examples of the main governance initiatives to date, which have resulted in various sets of guidelines and principles. So far, these are all voluntary commitments. The ICRC, regarded as the guardian of IHL, has called for new binding international rules. It advocates prohibiting some kinds of systems—including those with effects that cannot be sufficiently understood, predicted, and explained, and those designed or used to apply force against persons directly—and "strict restrictions" on other kinds.

TABLE 1. Governance initiatives on AIS in military contexts

Level	Examples
Global	(2023). General Assembly First Committee Resolution: Lethal Autonomous Weapons Systems (UN Document No. A/C.1/78/L.56) (2019) 11 Guiding Principles, Convention on Certain Conventional Weapons (CCW)
Cross-Regional	(2023) REAIM Call to Action (2023) Political Declaration on Responsible of AI (US-led)
Regional	(2024) Freetown Communiqué (Economic Community of West African States) (2023) Belen Communique (Latin America and the Caribbean) (2023) CARICOM Declaration (2021) NATO Principles of Responsible Use of AI in Defense (2021) European Parliament Resolution on Artificial Intelligence incl military use
State	(2022) Ambitious, Safe, Responsible: Our Approach to the Delivery of AI-Enabled Capability in Defense (UK) (2022). Provisions contained in the Arms Control and Disarmament Strategy (Switzerland) (2020) AI Principles: Recommendations on the Ethical Use of AI (US) (2020) A Method for Ethical AI in Defense (Australia) (2019) AI For Defense (France)

In October 2023, the First Committee of the UN General Assembly adopted a resolution stressing "the urgent need for the international community to address the challenges and concerns raised by autonomous weapons systems." The Secretary General has likewise recommended, "a legally binding instrument to prohibit lethal autonomous weapon systems that function without human control or oversight, and which cannot be used in compliance with international humanitarian law, and to regulate all other types of autonomous weapon systems." At the same time, there is a recognition that being compliant with IHL may not be sufficient to demonstrate that a system is responsibly deployed throughout its lifecycle.

1.3. WHY A LIFECYCLE FRAMEWORK

Considering the full lifecycle of systems—from development to procurement, testing, deployment, use, sustainment, and decommissioning—the framework offers a granular view of stakeholder involvement, clarifying when and how challenges should be addressed, and identifying who should be responsible and accountable at each lifecycle stage. It can also allow leaders to consider each project holistically, and gain insights at every stage about potential problems, conflicting guidance, or legal jeopardy.

Research Group members agreed that most of the challenges identified in the 2021 document, although seemingly technical in nature, remain human-centric despite the autonomous nature of the systems under discussion. Similarly, there was agreement that most of the challenges are not one-off problems to be solved but instead need to be addressed by humans repeatedly—when a system is first considered, then developed, tested, deployed, sustained, and retired.

Most discussions around AIS in defense settings tend to focus on two aspects: First, discussing what happens if the system functions in unintended or unpredictable ways, and how to assign responsibility and accountability for failure. Second, discussing what IHL permits, prohibits, and requires. Both discussions require addressing human decisions throughout the entire lifecycle of the system, including well before it is even considered for deployment.

Everyone associated with the deployment of an AIS for military purposes needs to have a common understanding of its capabilities and limitations: how it works, what criteria will trigger the use of force, the risks and potential consequences of failures, and how geopolitical context could affect its impact. Reaching a common understanding is not always easy given the professional diversity of stakeholders involved, such as lawyers, ethicists, and technologists, an issue further explored in 2.2.

A system cannot be understood in isolation of the human personnel it will interact with for a given task or its intended environment of use. Do the humans involved understand what the system will do? Are there well-defined, well-understood, and well-motivated transition points and feedback loops between human and machine activities, actions, and decisions? Is the environment dynamic or static, observable or not, communication-denied or not?

This lifecycle framework can better enable all involved to define and design around intent as early as possible and maintain that throughout the lifecycle. This should help human operators understand the context within which they are working. It should also lead to revisiting aspects of the project before use if the understanding of context and intent has changed.

In other words, the lifecycle framework makes clear that a system should not be used unless it has documented redundancies; clear documentation about contexts in which it can reasonably be deployed; an intuitive user interface; processes that take into account human psychology and cognitive limits; and people who understand how the technology works. Addressing these and other specific issues identified in the framework offers a way to operationalize general principles such as being reliable, trustworthy, and responsible.

This work builds on previous thinking about an AIS lifecycle and the role of humans during stages of development and use. The AIS lifecycle developed by the Research Group is more comprehensive, fine-grained, and modular than previous models. The greater the potential for autonomy in a military system, the more critical it becomes to consider human-centric design principles as early and as often as possible.

The work of the Research Group is a step towards operationalizing and clarifying such a human-centric approach to AIS. Often, it raises challenges and questions rather than providing definite answers. These answers could be spelled out in a formal standards process around AIS in the military context. Such an approach could build on the proven track record of industry standards as consensus documents drafted by expert teams to provide operational-level guidance and a scientific basis for regulation.

2. INSIGHTS FROM THE PROCESS

2.1. WORKING BACKWARD FROM AN EXAMPLE SCENARIO

The Research Group developed the lifecycle framework over two years of group and individual meetings. They approached the task by first developing an example scenario. They then worked backward from this scenario to explore, test, and validate stages in the lifecycle as they developed it. This document is available as a separate record online and shares notes from discussions about the scenario and the framework to provide insight into the thought processes involved.[3] An abridged version of the fictional scenario is as follows:

> As food insecurity increases worldwide, an international armed conflict results in threats to ships carrying grain. Researchers in academia and industry develop an uncrewed underwater vehicle (UUV) with capabilities to defend grain ships. It is called the Food Defender (FD). When a national military deploys the FD, one of the following occurs:
>
> 1) It attacks and sinks an adversarial ship while ignoring friendly ships.
> 2) It fails to attack an adversarial ship, and, as a result, a grain ship is attacked.
> 3) It is hacked by an adversarial system and disabled or used to attack a grain ship.
> 4) It sinks a ship of refugees, having mistaken it for an adversarial ship.

The scenario was intended to describe a fairly straightforward combination of capabilities that exist today, based on unclassified information, but without resembling any existing system too closely. The hypothetical system was designed to limit the risk of civilian interaction so the Research Group's attention could be focused on other issues that could otherwise be under-addressed.

[3] For a full version of the example scenario, please see IEEE SA Research Group on Issues of AI and Autonomy in Defense Systems. (2024). Human Decision-Making Through the Lifecycle of Autonomous and Intelligent Systems in Defense Applications: Example Scenario with Comments, available at: https://doi.org/10.5281/zenodo.13837345

Working backward from the scenario proved to be a valuable approach. It enabled members of the group to remain focused. It clarified which issues need to be addressed at each stage of the lifecycle and how the same issues might need to be addressed at different stages by different people. This helped to differentiate stages that were more chronological from stages that needed to be addressed throughout most of the lifecycle. The result was a well-rounded framework that can aid in considering the potential impacts of actions at later stages, problems that might arise, and questions of who should be held responsible and accountable.

Four general lessons emerged from the process of developing the framework.

1) Addressing questions at appropriate stages: Many conversations about AIS in defense contexts focus on the moment when the system is used. However, the group found that questions arising at that stage tended to lack focus and specificity. Considering the same questions at earlier stages provided a better sense of how to address them before the system was deployed, identify who should address them, and establish a chain of responsibility and accountability for when something went wrong. This could be addressed by following an ethics-by-design approach.

2) Awareness of trade-offs and risks in complexity: Autonomous and AI-enabled capabilities can appear to simplify the operation of a system at the point of use. However, the development and oversight of such capabilities can—and, at least for the foreseeable future, will—add complexity at an earlier stage. Militaries and governments considering AIS for military purposes need to gain a clear understanding of the costs, benefits, tradeoffs, and risks of developing or adding autonomous and AI-enabled capabilities to a system, and be prepared to make greater investments in people as well as technology.

3) Involvement of technical people from civilian sectors: It was challenging to find people with technical backgrounds to be part of the Research Group. Many who worked in related fields, such as aviation, autonomous vehicles, or medical diagnostic software, suggested that their work was unrelated to defense. However, there is uncertainty about which technologies might be available and weaponized or militarized in the future.

Technical experts can help non-technical people—such as those with legal and ethical backgrounds, policymakers, and military leadership—understand how technology works, its capabilities and limitations, and what questions to ask. Advancements in controls, standards, and policies for AIS can be achieved more effectively if technical and non-technical people are involved in the conversations. Different stages of the lifecycle tend to have a preponderance of one type of expert—e.g., technical experts in research and development and procurement and acquisition—rather than the diversity of experts needed for a more

sufficient awareness of the multiple concerns AIS raises.

4) Importance of documentation: Technical and legal documentation at each stage of the lifecycle is critical to give context and understanding to people interacting with a system at later stages of the lifecycle. It is the most effective means for providing commanders and operators with the information they need to understand what the system can and cannot do.

Documentation can also help verify that testing regularly confirms the system is functioning as intended and could be used in compliance with IHL and that human users understand how and when to use the system. It helps to establish responsibility and accountability throughout all stages of the lifecycle and during transitions between stages. For example, legal reviews conducted at different stages of the lifecycle should be clearly documented and known to legal advisers and commanders on a particular mission.

2.2. WHERE THE GROUP AGREED AND DISAGREED

Observations about the Research Group's experience working on issues related to AIS in military contexts may help other multistakeholder groups of experts to better collaborate and overcome misunderstandings. Participants had different ideas and understandings about the types of technologies to be considered and their current capabilities.

For example, when developing the scenario, military experts described a UUV they thought would be legal. Legal experts assessed aspects of it to be in violation of IHL. They offered suggestions for improvement, but roboticists pointed out technical limitations. Only after extensive and nuanced discussions could all members agree on the scenario. Challenges such as these highlight how critical it is to bring together experts with different backgrounds and expertise to consider the full complexity of AIS in defense contexts.

Different experts have different understandings of terminology and capabilities. When the scenario was tested with wider groups of experts, about half agreed that a system like the FD was realistic to imagine within a timeframe of around five years. A quarter believed that the technology was further off, while the other quarter said it already existed.

Given how contentious discussions about AIS in military contexts can be, another significant finding was how much agreement there was among the group on a variety of issues. The Research Group comprised people from very different backgrounds. The disagreements centered on definitional issues, limits of acceptability for the use

of AIS in military contexts, and the question of whether or not such systems should be held to a higher standard than humans. These areas of disagreement were relatively limited and outweighed by many points of agreement, including the following:

- There is value in considering a risk-based framework for decisions about the development, deployment, and employment of AIS in defense contexts.
- It is beneficial to categorize AIS into risk categories, accounting for factors such as predictability, type of target, duration, and environment of use. For example, all agreed that AIS for nuclear weapons would fall into an "unacceptable" risk category.
- General purpose robots, especially those designed for civilian or non-military use, should not be weaponized. (Although members of this Research Group agreed on this point, many reviewers expressed uncertainty and suggested the statement was too broad.)
- A system cannot be acceptable for use unless commanders and operators have sufficient contextual understanding of how it works and what it can and cannot do. A system would, for example, be unacceptable if the user cannot foresee whether the effect of the attack may be unlawful.
- When considering what to allow or regulate, it may be more helpful to think about autonomous and AI-enabled capabilities—e.g., targeting—rather than try to broadly define and regulate AIS themselves.
- Systems should not be able to learn and update in the field without proper testing, evaluation, verification, and validation (TEVV), legal review, and addressing how the update might affect the contextual understanding of the human operators. Because of the substantial technical challenges to such fast and flexible TEVV, this means AIS should not be able to learn and update in the field.
- AIS should be reliable, predictable, and understandable in their functioning and effects. The use of a military AIS that functions erratically could be indiscriminate and, therefore, in potential violation of IHL.

2.2.1. A NOTE ON GENERATIVE AI

Rapid progress in generative AI came to public attention only after the example scenario had been developed. Widespread publicity of its associated problems demonstrates the need for much more transparency, maturity, understandability, and reliability before it should be considered for use in highly consequential applications. This is especially important given the expected increase in the employment of AI agents—software entities that perform some tasks autonomously—and the potential use of such agents in future AIS.

3. INTRODUCING THE LIFECYCLE FRAMEWORK

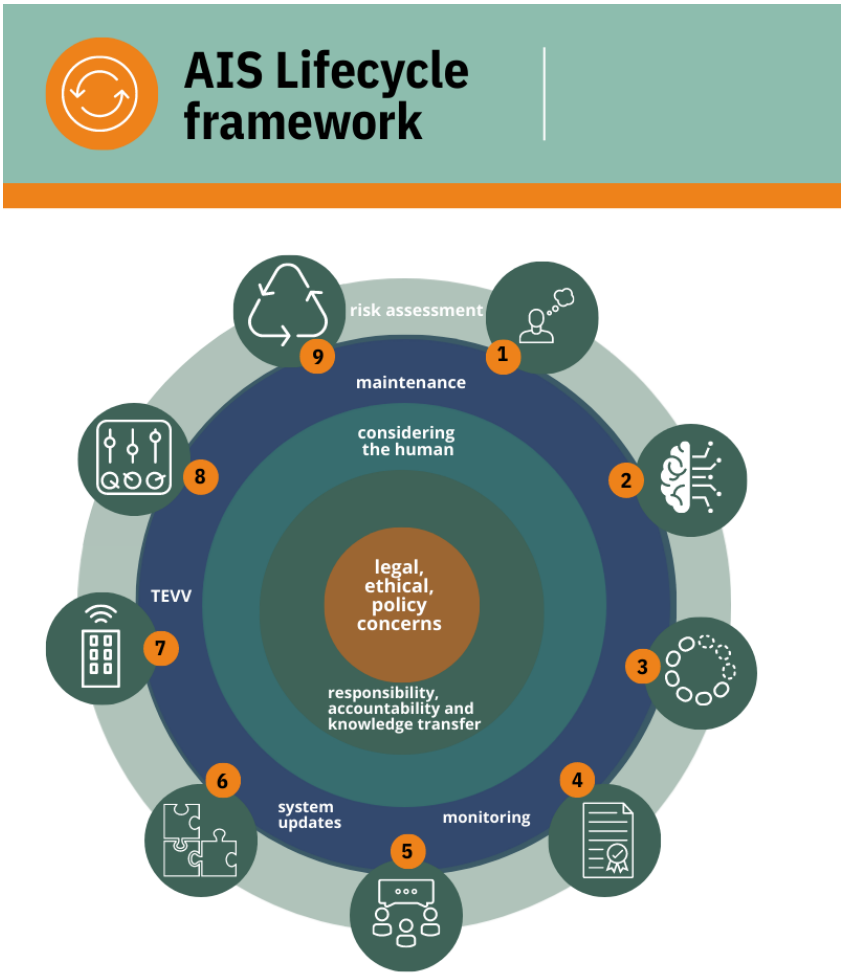
3.1. LIFECYCLE STAGES AND ONGOING ACTIVITIES

Members of the Research Group concluded that there are nine "lifecycle stages" in which human decision-makers play key roles (see FIGURE 1). Between these, there is likely to be a transfer of responsibility and accountability to new human decision-makers. Additionally, the group identified five "ongoing activities" in which human decision-making elements occur repeatedly throughout the lifecycle.

The nine lifecycle stages are as follows:

1) **Before AIS development**
   - General legal, ethical, and related technical concerns identified and addressed
   - Rationale for military development and use of AIS, formulation of system requirements, and considering the role of researchers
2) **Research and development**
3) **Procurement and acquisition**
4) **Testing, evaluation, verification, and validation (TEVV)**
5) **Considering the human: Education, training, and human-system integration**
6) **Political and strategic considerations**
7) **Operational-level command and control**
8) **Tactical employment**
9) **Review, reuse, and/or retire**

FIGURE 1 **Lifecycle stages**



1 **Before AIS development**
Identify and address general political, ethical, legal, and technical concerns. Define rationale for military use of AIS and formulate system requirements

2 **Research & Development**
Identify and address concerns that arise during system design, including through assurance, testing, and evaluation

3 **Procurement & Acquisition**
Verify that AIS fits the requesting organization's requirements, including through assurance, testing, and evaluation

4 **Test, Evaluation, Verification, Validation**
Various procedures to verify that system can perform its intended functions, i.e. meets system requirements

5 **Considering the human: Education, training, and human-system integration**
Design education, training, and testing with the various political decision-makers, commanders, and users who will interact with AIS in mind

6 **Political and strategic considerations**
Political leaders need to understand how AIS are different to "traditional" systems, including operating parameters, costs-benefits, potential failures etc

7 **Operational level command & control**
Processes by which operational-level commanders decide on AIS use, oversee operations, maintain awareness of AIS' operating areas, and maintain authority, direction, and control of AIS

8 **Tactical employment**
Tactical units require detailed understanding of AIS' performance, limitations, failure modes, risk mitigation measures etc

9 **Review, reuse and/or retire**
Review to determine if the system should be reused as is, updated or modified for future use, or retired from use

Some of these lifecycle stages are necessarily sequential—for example, a system must be developed before it can be deployed. Many, however—especially during the first half of the lifecycle—may overlap significantly and start or end at around the same time. For example, although the research and development stage is listed before procurement and acquisition, elements of the latter may occur before the former.

These stages may also be repeated many times after a system has been deployed and when it receives new hardware or software updates. The expectation is that mini-cycles of the stages will occur repeatedly throughout the lifecycle of a system. It is recommended that the nine stages be considered more as guidelines on how to think about what is happening with a system at any given point and who is responsible.

The five ongoing activities are as follows:

1) **Evaluation of legal, ethical, and policy concerns**
2) **Responsibility, accountability, and knowledge transfers**
3) **Considering the human: training, education, and human-system integration**
4) **TEVV, monitoring, hardware system or software updates and interoperability, maintenance**
5) **Risk assessments**

These activities need to take place at every stage so that humans can remain accountable for the legal and ethical operation of a system. Each of the activities will need to occur at most stages, and some will occur at all stages.

Two of the ongoing activities (3 and 4) are also lifecycle stages (4 and 5). They are lifecycle stages because they have discrete beginnings—for example, both will be more relevant after the system has been developed. They are ongoing activities because they must occur regularly throughout the rest of the lifecycle and during any mini-cycles. Members of the Research Group felt that these stages and activities were sufficiently important to justify the redundancy of including them in both lists.

3.2. CROSS-CUTTING FACTORS TO BEAR IN MIND

The next two sections of this document set out in more detail the Research Group's thinking behind the ongoing activities and lifecycle stages. Members of the Research Group suggest the following six factors to bear in mind when reading them:

1) **Human-centricity:** The lifecycle framework highlights the role of humans. Each stage and ongoing activity is designed to identify the relevant decision-makers, what their role at each stage is, and the

knowledge they need from previous decision-makers.

2) **Redundancy:** Many aspects of development, testing, and decision-making may need to be revisited throughout different stages of the lifecycle to verify that a system is working correctly and to be clear about who is responsible and accountable.

3) **Non-exhaustiveness:** The framework is intended as a starting point for discussions. It can be combined with other guidelines. It is not intended to include all questions and issues that policymakers, militaries, and weapons manufacturers/ developers should consider.

4) **Autonomy:** The framework focuses on aspects of autonomous systems that differ from "traditional" systems. As such, many issues are not addressed here because these apply to weapon systems more broadly rather than AIS specifically, including standing objections to the use of military force for lethal purposes. This focus is intended to inform discussions about what new policies, regulations, or laws may be needed.

5) **Applicability:** The framework applies to all militaries, while acknowledging that some points may be more applicable to some militaries than others given differences in technical expertise, national laws and regulations, and ethical requirements and standards.

6) **Flexibility:** Although each stage in the lifecycle must occur, they may not necessarily occur in the order described. The order will differ depending on factors such as military practices, research practices, how militaries partner with private companies, and political and military leadership.

4. FIVE ONGOING ACTIVITIES

The ongoing activities address the most important aspects of human responsibility and accountability and add redundancy to technical development and human processes related to AIS in military settings. At the beginning of a potential project, how these ongoing activities will be undertaken and by whom throughout should be defined.

4.1. EVALUATION OF LEGAL, ETHICAL, AND POLICY CONCERNS

Given the evolving nature of AI, autonomous capabilities, and battlefields, a number of legal, ethical, and policy concerns need to be addressed at every stage of a system's lifecycle. Legal experts should be involved from the conception of a system rather than waiting until it is in development to consider its legality. Compliance with the provisions of this document does not guarantee compliance with applicable legal and regulatory requirements. Legal input needs to be reassessed and verified at every stage thereafter, especially but not only if there are material changes to the algorithms integrated into the system. Many have argued that algorithms or models have not yet been shown to be useable in compliance with IHL in the absence of human verification and authorization. Capabilities that enable an AIS to apply force should be under stricter scrutiny to determine if it can even be developed, let alone deployed.

Different states have slightly different legal obligations. States party to Additional Protocol I to the Geneva Conventions are under an obligation to determine "in the study, development, acquisition or adoption of a new weapon, means or method of warfare" whether its use would be prohibited by the Protocol or any other applicable rule of international law. It has been argued that states not party to this protocol are also obligated to undertake legal reviews of weapons as a matter of customary international law and under a state's obligation to "ensure respect" for IHL. Independent of the customary international law status of Article 36 of Additional Protocol I, any military or weapons developer should regularly review legal, ethical, and policy concerns at every stage of the lifecycle.

4.2. RESPONSIBILITY, ACCOUNTABILITY, AND KNOWLEDGE TRANSFERS

Responsibility and accountability need to be transferred across the lifecycle stages. For this to happen, knowledge about the system and the humans involved also needs to be documented and transferred.

Responsibility and accountability: There must be a clear chain of responsibility throughout the lifecycle and within the stages. At any stage of the lifecycle, it should be possible to look back to see who was responsible for each stage or project within a stage. It needs to be clear how, when, and to whom their responsibility was passed during transitions between stages. Questions to consider include the following:

- What documentation is necessary to clearly define responsibility and accountability at and throughout each stage?
- What documentation is necessary to verify responsibility and accountability have clearly been transferred from one person or group to the next?
- Who is responsible for documenting responsibility across the lifecycle and at each stage, and what is the process?

Knowledge: A through-line of knowledge should be developed and maintained to verify that a system is legal, will behave predictably, and will be employed as expected. This requires clear communication about capabilities, limitations, and risks, often among people with disparate roles and areas of expertise—for example, engineers and military commanders. Knowledge transfer can help everyone along the line better understand the results of TEVV, risk assessments, and legal, ethical, and policy concerns and solutions. This adds redundancy for responsibility and accountability, so similar questions apply:

- What documentation is necessary to verify responsibility and accountability for knowledge transfers within and between stages?
- What documentation is necessary to identify human decision-makers clearly?
- What barriers to communication between disparate parties need to be addressed to enable successful knowledge transfer?
- Who is responsible for documenting knowledge transfers across the lifecycle and at each stage, and what is the process for this?

4.3. CONSIDERING THE HUMAN: TRAINING, EDUCATION, AND HUMAN-SYSTEM INTEGRATION

Humans are involved at all stages, including developers, designers, policymakers, commanders, soldiers, and operators. They need varying levels of understanding about what the system is, what it can do, and how the operators and system will interact in the intended environment of use. The range of experience will increase if systems are used by humans across states and cultures. Interface design and interoperability—when various systems integrating autonomy and AI are used conjointly—are key concerns.

Human-system integration

- Many problems in the past arose because designers did not understand or predict how operators would use—or might misuse—a system. Training, though critical, does not necessarily mitigate this issue. Human-system integration needs to be part of the design and all stages after.
- Humans have a tendency toward automation bias and other cognitive limitations, such as finding it difficult to maintain focus for extended time frames, monitor repetitive tasks, or stay engaged with secondary tasks as workload increases. These need to be addressed in a system's design and considered and monitored through all stages.

Education

- All humans involved need to have a relevant understanding of the system's capabilities, limitations, purposes of use, risk factors, etc. What is relevant will differ for different human roles and systems.
- Their understanding needs to be updated whenever relevant updates are made to the system. What is relevant will need to be defined for each system and updated as necessary.
- Decision makers need to be made aware of a technology's capabilities and limitations before making decisions about purchasing or building systems, and again before fielding systems.

Training

- As with other weapon systems, those who authorize and make use of a military AIS (i.e., operators and commanders) will need to be trained in how the system works.

- Those who authorize and make use of the system may need new training if the scenario for using it changes.
- Those who authorize and make use of the system will likely need new training if its functioning changes, such as through software updates.

Intent

Intent will look different at different stages. Operational intent to engage a particular target will be different from the general intent of engaging certain types of targets. Human intent needs to be understood at all stages, not just at the final stage. Political intent needs to be clear across the entirety of the lifecycle, especially in particular contexts of use.

Realistic conditions of operation

When humans actually operate AIS, the reality differs from design and planning contexts. For example, they may be strained by combat conditions, shortfalls in personnel, lack of rest, or larger-than-expected workloads.

4.4. TEVV, MONITORING, HARDWARE SYSTEM OR SOFTWARE UPDATES AND INTEROPERABILITY, MAINTENANCE

These four elements are grouped together because each can have a causal effect on the others. For example, if someone monitoring a system identifies a problem, that system will likely need some sort of maintenance or updates, which will then require a new round of TEVV. Alternatively, routine maintenance may require system or software updates which will again require a new round of TEVV and may also require higher levels of monitoring for a time to verify the system "acts" appropriately.

4.4.1. TESTING, EVALUATION, VERIFICATION, AND VALIDATION (TEVV)

TEVV happens as a system is developed and during or after purchase—in the lifecycle, across stage 2 (research and development), stage 3 (procurement and acquisition), and stage 5 (education, training, and human-system integration). Also, some level of TEVV or other type of confirmation needs to occur whenever the system receives new data that causes it to update (e.g., retraining in a field development setup) or an actual online update.

4.4.2. MONITORING

A defined person or set of people will need to be responsible for overseeing the system when it is in use or when it is brought online for use and has software running or updating in the background. Even if monitoring the system becomes in itself an automated process, someone still needs to be clearly responsible for verifying that adequate monitoring is taking place.

4.4.3. HARDWARE SYSTEM OR SOFTWARE UPDATES AND INTEROPERABILITY

For general maintenance and security reasons, a system's software, data, and data training or recalibration will need to be updated regularly. This will require additional monitoring and TEVV and may trigger new legal reviews if the update affects the effects of the systems significantly.

4.4.4. MAINTENANCE

This can include routine hardware or software updates, or fixing a system after it was damaged in battle, training, or testing.

4.5. RISK ASSESSMENTS

A system might pose a risk for various reasons, such as it did not function as intended; it was used incorrectly; it was poorly trained, or trained on incomplete data that do not span operational uses; it was used in the wrong situation; the user did not understand its limitations; the developers misunderstood how the system would be used; or it was hacked or otherwise adversarially manipulated. The use of a system might also pose political risks, risks of proliferation, or risks of escalating a situation. Risk assessments must occur regularly throughout every stage of the lifecycle. They will depend on the system, the expected operational context, and the stage.

Good starting points for risk assessments related to AI include the Organization for Economic Co-operation and Development (OECD) High-Level AI Risk Management Framework, the National Institute of Standards and Technology (NIST) AI Risk Management Framework, and the framework developed by the National Academies of Sciences, Engineering and Medicine. However, these will need to be further developed to apply to military systems, which inherently carry more risk than non-weaponized AI and involve different trade-offs.

5. NINE STAGES OF THE LIFECYCLE FRAMEWORK

The nine sections of this chapter provide detailed descriptions of concerns and challenges that came up during group discussions on each stage of the lifecycle framework. There is some unavoidable overlap with the preceding text that summarized issues cutting across the nine stages. The aim is not to be comprehensive or exhaustive, but to highlight issues that need to be addressed and provide a framework for identifying other issues that will arise as individual AIS are developed for military uses.

5.1. BEFORE DEVELOPMENT

5.1.1. GENERAL LEGAL, ETHICAL, AND RELATED TECHNICAL CONCERNS IDENTIFIED AND ADDRESSED

Much of the focus in international debates on AIS in defense centers on the potential for human control to be exercised at the point of deploying the technology. However, many people are involved across the entire lifecycle of such systems, including before their development has even started. This stage includes the various governance discussions happening currently on the ethics and regulatory needs for AIS in military contexts at global, cross-regional, and regional levels. Key issues include the following:

1) International law, including IHL, applies to AIS. However, it is not always clear how it does so, and some participants in governance discussions have suggested that new, specific binding rules may be necessary. Questions that might help address legal disagreements include the following:

   - What does it mean to "automate the kill chain," and what regulations are necessary?
   - What rules should exist regarding the use of a system that cannot communicate with responsible parties? Or should constant oversight/communication be required?
   - Weapons tend to be deployed faster during times of war. Can regulations allow for sufficient time to verify that AI and autonomous capabilities function as intended and that humans have a sufficient understanding of how to deploy the system?
   - Is there potential now to reach agreement on types or uses of AIS that should never be allowed in military settings, because everyone agrees they are unethical or already illegal under IHL? For example, the Research Group agreed on not weaponizing general-purpose robots, not using

machine learning in nuclear weapons control systems, not using unpredictable systems, and not allowing systems to learn and update in real-time during battle.

- Should humans use AIS for military purposes only if it can provide a higher degree of compliance with e.g., IHL than humans using traditional, non-automated systems in the same scenario? Discussions in the Research Group suggested this may be an especially challenging question to address. Because AIS and humans can fail in different ways, making a direct comparison is difficult. It also risks falsely equating the ways in which humans and systems process information and reach decisions. Even if it is agreed in principle that a system must meet or exceed human expectations to justify its use, what that looks like is not clear.

- Where and how do reviews of the lawfulness of weapons, means, and methods of warfare—whether conducted pursuant to Article 36 of Additional Protocol I to the Geneva Conventions, or otherwise—fit in? For weapon systems, making an Article 36 review mandatory at all design reviews was seen as a potential starting point. Article 36 reviews must consider the ability of the system to be used in compliance with IHL and other applicable legal frameworks.

2) The terminology challenges around AIS in defense settings show the need for a greater understanding of what makes AIS different from traditional weapon systems.

- Is it clear what problem is being solved by using autonomous and/or AI capabilities? Are these capabilities the best solution for that problem? Is it clear what the desired function is for the system? Is it clear that autonomy or AI are the best solutions to perform that function? Would this lead to skills obsolescence, which may be a disadvantage if the system cannot function? Militaries are often not necessarily looking for the “best” solution, but a solution that can be adapted quickly for military use. A technical specification of what a technology is supposed to do could be a first line of defense against the rapid deployment of systems with capabilities that are not understood by their developers.

- Autonomous and AI capabilities pose different challenges and have different requirements from previous weapons. This does not mean AIS represents a discrete class of weapons; rather, these capabilities can be added to all kinds of weapons. These distinctions should be identified now, to as reasonable an extent as possible, so that groups can define rules and regulations specifically for systems integrating autonomy and AI.

- In some cases, already-deployed weapons technology should be revisited to determine if it should be classified as including autonomy or AI functionality. This would not necessarily mean that the system can no longer be used, or that new regulations are required. Rather, the system may require a new, objective assessment of the capabilities in use. Looking at existing and near-term systems will provide a deeper understanding of challenges throughout the lifecycle and human involvement at every stage. Existing systems may also provide guidance for understanding future technology that might be deployed and what regulatory frameworks may be necessary.

- It needs to be clear, from both a responsibility and a technical standpoint, when the human is in control of the system and how that transfer occurs. This needs to take into account that control may shift back and forth between the human and the system during certain missions. Such handoffs are a technical problem that needs to be solved in advance. However, this is also a question of user interface, training, and contextual understanding that will be dependent on—and need to be addressed for—each system and each mission.

Setting out general requirements:

1) Assessments to consider the trade-offs in developing an AIS in a military context—between benefits, risks, costs, and opportunities—should go beyond standard weapons assessments. They should include consideration of how the system might impact mitigating or exacerbating international AI arms races, and how levels of transparency about the capabilities and limitations of the system might change that impact. They should also assess if the technology increases complexity and/or cost across the lifecycle. For example, some assume that integrating autonomy or AI means systems will be cheaper or fewer humans will be involved because they focus only on the deployment stages. When assessing the full lifecycle, the opposite is often the case. Some stages—for example, early development and testing—may increase the overall cost and involve more humans.

2) Before any potential system development, it must be identified who should be involved throughout the lifecycle. This includes figuring out how to consider and address ethical, moral, and legal considerations throughout the lifecycle.

3) Where does the data used to develop autonomy and/or AI in weapon systems come from? Who has ownership of the training data? What types of data are used, and what is their quality? There are, for example, risks of using human-operated drone strikes as a baseline. Where will the data be stored? What

would happen if the data were hijacked? Is existing law sufficient to cover data use across multiple countries? Is existing law sufficient to protect civilians from military use of personal data? Who vets the data, and what is the process to make sure it is updated and synchronized with minimum latency? Are the current vetting systems adequate for continued optimal performance and security? If not, how will new vetting systems be developed and assured? How is new data processed for ingestion, and how are models based on the new data assured before deployment?

4) More research needs to go into understanding and defining what a clear set of requirements for intended use would look like for a potential military AIS. Once that is established, it can be used to develop documentation for design decisions, trade-offs for the system versus more traditional weapon systems, plans for addressing human oversight throughout the lifecycle, etc.

5) Problems with respect to human control or oversight at the deployment stages may arise because they were not addressed at earlier stages, including research and development. These problems need to be understood before potential system development, though there is no guarantee that they will have a solution. It must also be considered what responsibility means at the research and development stage, and how the role of the developer needs to differ for systems with autonomy and/or AI compared to traditional weapons.

6) How can greater communication be fostered between developers, policymakers, and end-users? Language should be both technically meaningful and understandable to the humans who will be applying the system.

7) Between commercial and military needs and uses of autonomy and AI, there is increasing guidance for testing, evaluation, verification, and validation. However, much more work is necessary to meet the requirements of international law, including IHL and other military needs.

8) How can political leaders be sure they have the information they need about capabilities and limitations associated with potential systems?

9) To what extent can the technical challenges of AIS in military settings be addressed through general guidelines, standards, or regulations, as opposed to those specific to the development of individual weapons? How can general guidelines, standards, or regulations be addressed now?

Requirements for humans throughout the lifecycle:

1) Intent matters. That includes the intent of human operators regarding what they expect the system to do. It also includes how well the human operators understand the capabilities and limitations of the system and the effects that will result in the intended circumstances of use, the original intent of the developers, whether developers understand how human operators will use the system, etc.

2) Infrastructure needs to exist to identify who among the humans involved across the lifecycle is accountable and responsible. Processes, documentation, record keeping of decision making, doctrine, logistics, and training all need to be developed. There should be some sort of auditable documentation trail so that if things go wrong, there can be oversight to identify where, when, and why and to record the learnings for the future.

3) Human training is just as important as the technology. Tailored training must occur for people throughout the lifecycle stages, including those involved in TEVV, procurement officers, military commanders, and those who will ultimately use the systems. Skills assessments may also include which skills need to be retained when technology fails.

4) How humans relate to machines is not well enough understood. More research is needed in human psychology, human factors engineering, and human-system integration, as well as technical research and development, to address problems such as: better predicting how humans will work with a system; how control will be shared between the human and system for a given action; how to provide for humans to have an appropriate balance of trust and skepticism of the system, and how to define that balance. The problem of automation bias also needs to be addressed, as it complicates the task of identifying when human control, involvement, or oversight of a system is meaningful. Automation bias is especially likely to happen when humans are overwhelmed by data, as can happen with AIS in military settings.

Cross-cutting concerns for civilian and military applications of autonomy and/or AI:

1) Many discussions about ethical AI focus on commercial uses, and avoid discussion of potential military use of the same technologies and capabilities. This means that militaries may not have access to best practices that already exist within the industry; questionable technology may still be developed for militaries, but with less oversight; and commercial applications may be repurposed for military use

without the developers' knowledge or consent. Naturally, ethical issues for commercial products will not be the same as those for weapon systems. However, there is significant overlap. Ethical adaptations for AI will be more robust if civilian technologists engage with militaries and vice versa. Discussions and decisions should include people from a variety of professional and disciplinary backgrounds, including the military, tech industry, technical academia, political academia, international law, and policymakers.

2) Just as new regulations are being considered for AI more broadly, new regulations may need to be considered for the use of advanced AI capabilities to enhance the performance of military systems. What can be learned from current examples of AI regulations that may be applicable to defense contexts?

3) Questions around AIS in defense settings are not just military concerns, nor are they all centered around IHL. Weapons will be used by other actors, and issues may fall under other bodies of international law—such as international human rights law (IHRL)—as well as national laws, or other legal requirements and norms. Even if guardrails are in place, there is no guarantee that a system will remain fit for purpose if it is used in different use cases than those for which it was originally designed. AI systems are vulnerable to proliferation, especially as the technology becomes increasingly available. The impacts of AI proliferation can be expected to spill over into AIS in military settings.

5.1.2. RATIONALE FOR MILITARY DEVELOPMENT AND USE OF AIS, FORMULATION OF SYSTEM REQUIREMENTS, AND CONSIDERING THE ROLE OF RESEARCHERS

While every military uses different processes for developing and acquiring new systems, the pattern is generally similar in all cases: determination that a new system is needed to meet defined or assigned missions and/or emerging threats; establishing hardware and software requirements for that system; initial system design; prototype development; testing and evaluation; decision to move to scaled production; funding approval and procurement; fielding; and long-term sustainment.

Before any of these processes begin for a potential AIS, the rationale for adopting it for military use should be established. This includes perceived benefits, opportunities, and risks, along with a general description of the process for establishing formal system requirements for both hardware and software. This process for defining requirements initiates the rest of the lifecycle. Issues to consider include the following:

1) How policy and legal reviews are integrated at various stages throughout the lifecycle, along with appropriate levels of oversight to verify the system is built to and tested against specified

requirements. What information needs to be secure, what needs to be communicated to all those working on and with systems, and how can issues or questions be escalated up the chain?

2) How requirements for AIS will differ from those for "traditional" hardware and software systems. Requesting organizations—a military service, for example—designers, developers, testers, end-users, and budget-approval authorities need to agree on initial requirements, and on changes to performance criteria and other metrics throughout the system's lifecycle. Hardware requirements may remain stable or even fixed throughout the lifecycle, but constant evolution in software requirements is inevitable, especially for AI models that must be updated regularly even after initial fielding. Organizations must anticipate the implications, including how that could affect hardware needs.

3) How a requesting organization must define and anticipate the requirements for a system to be developed. These requirements will be based on operational needs, which in turn are predicated on assigned missions and operational plans—though these may be vaguely defined at the point of the request—and emerging threats that jeopardize extant systems. Some systems may be designed and developed with various degrees of autonomy and automation that can be selected throughout a mission as required. Others may be designed as human-controlled, with one or more autonomous and/or automated subsystems. Some may be designed initially without autonomy, with fully- or semi-autonomous capabilities added after initial fielding, such as onboard automated, autonomous, or semi-autonomous weapon systems. This differing functionality will need to be considered from the start.

4) Funding for all systems is considered through a separate, yet linked, process involving some combination of planning, programming, budget, and execution. Funding timelines often diverge from the requirements process in ways that make it difficult to gain rapid funding approval for incremental updates to fielded systems, especially for software changes and maintenance demands.

5) System designers and developers design and develop hardware and software based on specified, approved requirements.

6) Autonomous capabilities may either be designed into hardware and software from the initial system design, or added later to an extant weapon system, assuming the existing hardware supports the requirements of the autonomous capability.

7) How training will be incorporated. Training is a critical aspect of effectively and safely operating any weapon system, and training provisions are always included in acquisition contracts as part of lifecycle management. Contracts typically specify requirements such as the number of people to be trained, the type and duration of training, and any certifications or qualifications necessary for operating the system. Contracts may also outline the responsibilities of the contractor and the government regarding system delivery. In some cases, the contract may provide for the contractor to provide initial training to military personnel upon delivery of the weapon system. Ongoing training, maintenance training, or specialized training may also be part of the contract. When not done by contractors, the acquiring organization(s) should provide the necessary documentation and training plans. Whether a system is "off the shelf" or specifically commissioned can have an impact on the relationship between manufacturers and the military.

8) Legal reviews of AIS design requirements for military uses will include consideration of IHL/LOAC and other relevant legal frameworks, including those posited in Article 36 of Additional Protocol I to the Geneva Conventions.

9) In many cases, the potential development of AIS for military uses will involve different countries, militaries and companies. How will all of these groups work together and what are baseline ethics? The humans who will be accountable throughout should be identified; all necessary legal requirements of partner countries should be met; and technical standards should be consistent throughout all aspects of the project.

10) How to revise or retire the system if—during development or once built—it does not meet military standards or legal requirements. What is the process for preventing the proliferation of systems that were built but not fielded?

11) All challenges in the 2021 document should be reviewed at this stage. Some will need to be addressed here, and others at later stages of the lifecycle depending on the system being considered.

5.2. RESEARCH AND DEVELOPMENT

Autonomous capabilities that enable weapon systems are not necessarily separable from those that support non-weapon systems. For example, obstacle avoidance differs only in what is avoided; targeting a weapon is the same underlying technical challenge as taking a picture; and when an autonomous system attempts to position itself relative to a specific object, it does not matter why. Because this technology is fundamentally dual-use, an AIS may not always follow the same research and development path militaries are used to.

The research and development process for AIS may not begin with traditional defense contractors, but could start in academic or other industry labs. Initial research and development may not be done with weaponization in mind. While data-centric AI-supported systems may be difficult to re-purpose, components of AI systems developed for one task can often be repurposed for other tasks with minimal reconfiguration. Components of potential military AIS can be modified or updated independently of one another.

Improved components do not necessarily translate to improved performance of a system as a whole, which relies on the integration of components and their joint interaction with the environment. The software component of systems may be easier to update or modify than hardware components, which may itself create challenges for data processing needs. The implications of the dynamic and integrated nature of potential AIS for use and testing need to be considered at the research and development stages.

Individual AI components for perception, decision-making, and/or control that AIS in military contexts utilize may likely have been developed initially for non-military purposes. Research and development at government labs and contractors may largely be focused on repurposing, integrating, and testing the reliability of systems composed of AI modules.

To better address risks that arise during the research and development phase, implement effective mitigations, and enable system users to validate proper system operation and maintenance, the people developing and/or overseeing the development of systems need to:

- Understand the context within which the system is expected to be used.
- Understand what the system is intended to accomplish.
- Confirm they understand their role in the research and development process and that the process includes not only system design and development but also assessment and documentation of any aspects of the system that will affect its future operation.

Contextual understanding should include the following:

1) Identifying who is responsible for the system meeting IHL, IHRL, and other legal requirements. The developers themselves may not have detailed knowledge of legal frameworks, but should have a basic understanding and know who to go to for legal advice, as necessary to confirm regulatory requirements. Addressing the gap between the human language in which laws are expressed, and the specification languages used to characterize the behaviors of technologies, could be an opportunity for collaboration between engineers, legal experts, and policymakers.

2) Defining how accountability for components from different departments, companies, and states come together in a system so it can be developed, tested, and used appropriately.

3) Defining what the overall maintenance and ongoing monitoring plan is for each component of the system.

4) Determining how various direct and indirect stakeholders will interact with the system.

5) Determining how end users understand the output of the system.

6) Considering how releasing or publishing the technology, even if not initially intended for weapons, could affect weapons proliferation or enable misuse.

7) Determining how the security and integrity of the supply chain for the system can be validated.

8) Identifying when and how dual-use issues, along with any associated liability concerns, are identified and either addressed or flagged for users later in the lifecycle. Non-weapons developers should also consider this, if they seek to prevent their technologies or algorithms from being weaponized.

9) Identifying what processes are in place so that anyone throughout the lifecycle can identify and report problems with the system, without fear of repercussions, and knowing the problem will be taken seriously and addressed.

10) Establishing a process to stop the development or use of the system later in the lifecycle if problems are identified, and defining what type of problems should stop the development or use of the system.

11) Determining the processes, procedures, and fail-safes to allow the human user to take over the system if the mission changes, and when human intervention is not technically feasible. For example, this can include documenting conditions, especially temporal and geographic conditions, under which humans would not be able to intervene or override system behaviors.

12) Identifying how decision-making and target acquisition processes would change as a result of using the system, and whether it could be built to adjust to changes in rules of engagement, environments, mission sets, etc.

13) Recognizing that, due to concerns about proprietary information or security, many in the research and development stage will not know why they are working on some subtask—for example, labeling data.

14) Defining accountability for evaluating and mitigating ethics concerns.

15) Verifying that R&D funding requirements also have ethics built into the proposals and considering the issues pointed out in this document.

Understanding of system intent should include the following:

1) What the desired capabilities and acceptable limitations of the system are.

2) How the system can be deployed as a weapon and what the role of humans is in its operational use.

3) How well developers understand the problem they are trying to solve, and how much they do and can understand intent for the missions in which the system would be used.

Understanding of their role in the research and development process should include the following:

1) Defining who is responsible for the system meeting IHL, IHRL, and other legal requirements.

2) Identifying to what degree they are accountable for system performance even after system fielding.

3) Determining the roles they play in verifying the systems are used appropriately.

4) Verifying that the actual capabilities and limitations of the systems are compatible with the desired capabilities and acceptable limitations of the systems.

5) Documenting design assumptions that are likely to be broken during operational use by inexperienced or unknowledgeable users.

6) Implementing reasonable plans for TEVV and analyzing how the testing environment and data align with anticipated environments of use.

7) Identifying how the system satisfies various mission parameters.

8) Providing meaningful training and certification for personnel managing these systems.

9) Establishing the extent to which TEVV during research and development can be extended to real-world data and scenarios.

Other issues for developers and/or leadership to consider are as follows:

1) Performing comprehensive risk assessments, developing risk assessments for later stages, and creating early warning systems to monitor them. This includes risk assessments for academics and non-military developers to perform before deciding whether or not to share weapons-related use cases for their algorithms on open-source sharing websites.

2) Creating a process to analyze the function of components of an AIS to identify: a) how critical they are to overall system function; b) their modes of failure and their reliability; and, consequently, c) methods, such as run-time monitoring, by which overall system risk can be decreased.

3) Defining levels of rigor for algorithms or other control methods that determine the system's autonomous behaviors, while being aware of trade-offs. More rigorous algorithms can require more assumptions that are not always true in the real world, so getting a system to work may require balance in terms of rigor.

4) Determining core characteristics of a system, including whether a vulnerability was introduced into an AI model during training, the robustness of hardware, or the reliability of data.

5) Justifying why autonomous functionalities are preferred to “traditional” technologies or solutions. The benefits must be real, established, and observable, not hypothetical based on optimistic ideals.

6) Reporting how the system will function in various situations, how to limit and mitigate failures over the system’s lifetime, and how to verify an appropriate level of trust. This includes creating systems that are easy to test, characterize, and monitor, and reporting understandable limitations in behavior performance or current testing status. TEVV should form the basis of this report, with an understanding that additional TEVV is required for continued use or use in new scenarios.

7) Verifying that each component of the system is reliable and integrated appropriately, in order to help limit failures arising from the system when fielded.

8) Considering how human users will understand the system, limiting the use of the software in cases where failures would be more likely.

9) Creating systems that provide sufficient testing results or transparency for humans to appropriately calibrate their trust in the system. The ability of users to understand the reliability of the system will improve the ease of training.

Other considerations:

Developers of AIS or their components face the challenge of determining expected system performance in real-world domains. Claims about system behavior might be supported by a format such as assurance arguments, using a template that outlines what the system is expected to do, with specific claims and evidence. Claims about components or subsystems should be refined through research and development.

Even if a solid argument can be made that a system can achieve certain things under specified conditions, it will still be a challenge to determine how that translates to its potential use in new contexts. An assurance argument may be built around the physical structure of the platform, or functions that will not substantially change from one domain to another—for example, engine or brake reliability. It needs to be clear what parts of the argument people should focus on to understand if the system can adapt to a new domain. As the system is deployed to more domains, analysis of its performance may produce greater certainty as to conditions under which it will fail.

Developers can define failsafe behaviors that take effect when systems correctly identify certain states—for example, loss of communication or moving out of a certain region. This requires that: 1) undesirable states are known at design; 2) these states are detected during operation; and 3) behaviors created during the design phase are then executed and cause the system to exit that state. Due to contested information environments, designers may need to rely on their ability to identify which states are undesirable and behaviors to get out of those states. There may need to be a process by which human operators can designate states as undesirable and create corrective behaviors after fielding, as the use of the system changes.

Given legal and ethical obligations around AIS in military contexts, mechanisms and processes must be defined to assess acceptable risk quickly given the circumstances of operational use. For example, an object detector with a 0.9 probability of detecting a person would be unacceptable for use in determining whether a targeted area is clear of civilians before firing artillery. However, if an area has already been cleared of humans and artillery will fire in a few minutes, the same detector might reasonably be employed as a check on the human decision, if and only if procedures exist such that the human clearing the area does not rely solely on the algorithm.

5.3. PROCUREMENT AND ACQUISITION

While terminology differs, at some designated milestone, military and policy officials will approve the procurement of the system under development. Typically, though not exclusively, this will be a system combining hardware-software in an integrated way. Acquisition can also happen via diplomatic channels or even aid, though this is less common. Some of the concerns below may also occur during the first two stages.

Throughout the acquisition cycle, TEVV (described in the next section) will define criteria for moving forward. For example, a system may be required to conduct one successful test shot with a minimum reliability score to transition from one milestone to another. The organization tasked with testing and evaluating the system will communicate with the system's program office. Agility is required to effectively develop software to support a system's autonomy. Assessments of assurance will need to be continually performed and consistently communicated so messaging can stay consistent from the test team up to military leadership.

Issues to consider are as follows:

1) Procurement and acquisition personnel need to verify that a system is developed in accordance with the requesting organization's formally defined requirements, or capabilities definition. This includes

understanding the technical details associated with normal system performance and system failure modes.

2) Requirements for AIS will differ fundamentally from those for "traditional" hardware systems. Requesting organizations, designers, developers, testers, end-users, and budget-approval authorities need to agree on initial requirements, and on changes to performance criteria and other metrics throughout the system's lifecycle.

3) Procurement and acquisition personnel need to understand the differences between TEVV for AIS versus "traditional" hardware systems. Requirements and contracts need to reflect these differences.

4) Budgets need to account for the testing costs associated with modeling and simulation, digital twins, synthetic environments, laboratory testing, surrogate testing, testing against corruption and adversarial attacks, and so on. The potential for rapid change means AIS may be difficult to accurately budget for in comparison to legacy hardware.

5) Procurement and acquisition personnel should be integrated into the entire lifecycle, coordinating with designers, developers, testers, and end-users at all steps to provide transparency of approach and outcome.

6) In coordination with the requesting organization, designers, developers, testers, and end-users, procurement and acquisition personnel need to define and agree on criteria for entering and exiting each stage of the acquisition cycle.

7) Procurement and acquisition personnel shall verify that a system's requirements for training and certification are incorporated into contracts and/or accounted for in government-provided training plans. Training must include the IHL/LOAC implications and potential ramifications of both normal operating modes and system failure modes.

8) Procurement and acquisition personnel shall verify that contracts and/or government training programs account for follow-on training and certifications throughout the system's lifecycle, to account for software or hardware updates—including after a system has been fielded.

9) Data management and protection will be important, as will testing for integrity or resilience against corruption or adversarial attacks. Procurement and acquisition personnel shall verify that contracts

and/or government-issued documentation addresses requirements for AI algorithm performance, data management, and protection—including intellectual property protection—and TEVV throughout the lifecycle. There may be trade-offs between intellectual property protection and testability. Given that training datasets for AI models are usually proprietary, personnel may not be able to get precise information about what is included and independently verify documentation or statements from contractors.

10) Procurement and acquisition personnel should be familiar with the human-system integration aspects of a system and verify that contracts address initial and follow-on education, training, and certification requirements accordingly.

11) Legal reviews of system design requirements and subsequent acquisition contracts should include consideration of IHL/LOAC and other relevant legal frameworks. They should result in action as necessary to highlight potential problems and forestall potential violations.

12) Legal reviews of contracts for AIS in defense settings should address technology-sharing opportunities and constraints and/or export controls, which includes cross-checking with higher-level guidance up to and including national-level guidance.

13) Legal reviews of contracts should address how product liability, despite its limitations, might play a role in guaranteeing responsibility and accountability of persons manufacturing systems. This can include issues such as defective manufacturing, standards of fitness for intended purpose, system failures once fielded, and terminating further system development as necessary.

5.4. TESTING, EVALUATION, VERIFICATION, AND VALIDATION (TEVV)

Any potential system development should start with a set of requirements which, when met, enable the system to perform its intended functions. Once a system design is proposed, specifications are developed that describe what components must be used and subsystems must do in order for the system to satisfy the requirements. Components and subsystems can then be built or bought to meet the specifications, and assembled into a system which can be expected to satisfy the system requirements.

For this process to work, several types of verification need to occur. First, there needs to be a method for

verifying that the specifications will result in a system meeting requirements, for example of a legal or operational nature. This might include design reviews and simulations. Second, verification needs to confirm that the components and subsystems meet the technical specifications. This might be done by testing, or purchasing components that have some type of certification. Third, verification needs to confirm that the system, once built, meets the requirements. This can be done by testing the system, or by other means when the requirement does not need testing to be checked. It is always possible that the set of requirements is incomplete or has errors.

Validation is the process of checking that a system being designed and developed will perform its intended function. Tests could be created specifically for the system, though methodologies and equipment are often developed for a broad class of components or systems.

Evaluation includes evaluating the results of tests administered while checking requirements. It may also include evaluations performed for any system regardless of requirements, such as evaluations of human usability, environmental impact, interoperability, or cybersecurity.

Even after the system has been put into the field, it will typically undergo occasional testing and evaluation, for example, checking that software security is up to date, that emissions are within acceptable levels, or that safety features still function properly.

Complex and advanced autonomous capabilities complicate the TEVV process. In some instances, current procedures may not be sufficient. For example:

1) Meeting requirements such as “the system will never target a noncombatant” depend on being able to create specifications for the system, subsystems, and components that can be verified. For machine learning, data sets selected for testing generally serve as an implicit specification for the model. Anyone training machine learning models needs to be able to determine if the data selected for testing predicts performance in the environment in which, and at the task for which, the system will be deployed.

2) Testing of software typically involves identifying the type of input it will receive and verifying that it will then produce appropriate output. Possible inputs for AIS may vary so widely that only small regions of the input space could ever be tested. The range of input may also change from what was expected during development. Some type of monitoring during runtime may be needed to identify when input to the system is, in relevant ways, outside the bounds of input for which it has been tested.

3) Testing in simulation or with synthetic data may be used to understand what behaviors an AIS is likely to demonstrate in situations where physical testing is not practical. Simulated input must be of sufficient fidelity to elicit realistic performance. Synthetic data must be of sufficient quality to allow testers to predict the behavior of the component or system in the appropriate environment and for the appropriate task.

4) Testing for robustness to variation in environment, and other system inputs, will likely be needed to determine the risk associated with deploying a system to a new environment, or modifying and redeploying a system. This includes variation in commands provided by the user and interaction with other systems.

5) The system's user interface should be evaluated for usability, and to provide a suitable level of transparency and explanation to allow a user to invest an appropriate level of confidence in the system's ability to perform correctly.

6) Testing and evaluation should be part of an assurance process aimed at producing appropriate confidence in desirable system properties, such as safety, security, and reliability. That assurance should be documented in a form that justifies the confidence and communicates the conditions under which it is provided. The form of communication will be different for safety specialists, developers, purchasers, and users. Training for users should include conditions under which they may be confident in the system, and conditions under which confidence is not warranted. This may allow commanders to employ the systems in a manner that reduces risk. TEVV can never anticipate all conditions under which a system may operate, and at best can provide assurance under a set of conditions that are communicated to the user. It can also inform the user of risks, and methods of risk mitigation, when the system is used outside of those conditions. Test plans will need to be maintained and updated to cover the lifecycle of the system, including testing after potential fielding.

7) All systems should be expected to have potential defects, so a system needs a way of notifying the user when a defect exists and identifying the cause of the defect. A defect in the behavior of an AIS could be caused by changes in its environment which result in novel and untested input, as well as by manufacturing defects or software bugs. Methods for identifying defects during testing need to be developed. The process by which the system notifies the user of defects, or errors during operation, also needs to be tested. Since an AIS may operate out of contact with its user, its ability to detect defects, and execute failsafe behavior without human oversight, needs to be tested.

8) The process of updating system hardware or software is typically overseen by the manufacturers or developers, with the updated system going through TEVV before the updates are fielded. If AI machine learning models are part of the system, those models may be trained, tested, and deployed by the user, or at a rate that precludes using the typical TEVV process. Model applications may need to be categorized by potential risk, with different levels of test and verification rigor for different categories of applications. More broadly, the autonomous behaviors of the system may need to be categorized to determine what level of TEVV rigor needs to be applied before the system is used. Methods of TEVV for users who train models or change system behaviors in the field will need to be developed.

9) There is a lack of understanding of and practice in the TEVV of AIS in military contexts. Means are needed by which concerns about systems that pass TEVV can be identified and acted on to improve the TEVV process, and to inform users of potential risks. There are questions that remain unanswered, such as the following:

   - Should verification of a particular kind of system behavior be related to a representative mission in which that behavior might be used?
   - Do persons performing verification, testing, and evaluation need to understand the purpose for which the system will be deployed?
   - What new cybersecurity considerations are needed for AIS in military contexts? Should there be regular testing or recertification for cyber-related properties?
   - How should TEVV work for machine learning models or other system behaviors modified by the user?
   - How can the tradeoff between testability and security be handled? For example, testers may want to examine any part of code, which poses security risks.
   - How should testing be done between parties that do not trust one another and are unwilling to share software or data? For example, can testing be done between hostile countries to ascertain whether their systems will conform to international agreements?

- As the comprehensive nature of standard TEVV is infeasible for systems that use machine learning, what would a rigorous replacement look like? Is an internal causal analysis of machine learning systems needed?

- What should be acceptable error limits? While acceptable error limits must be part of requirement specifications and should be transparently disclosed to the user, evaluation of error limits will likely only be the basis for system induction. At this point, military users need to test that every requirement specification has been met.

5.5. CONSIDERING THE HUMAN: EDUCATION, TRAINING, AND HUMAN-SYSTEM INTEGRATION

The use of AIS will require additional training to that which members of the military already receive. Various political decision-makers, commanders, and users will interact with autonomous systems. Developers of any potential AIS need to have in mind that these people will have different training, technical expertise, and relationships with technology. Adequate human education and training are required at all levels, relative to the role or function that users and decision-makers occupy. Significant testing of the user interface is also required.

Human developers, commanders, and users need to understand how the system works, what will trigger the use of force, and how to use the system—what it can do, and its limitations. They need to be able to recognize when a system is not reacting as expected. This requires a sufficient understanding of context. Human education and training are necessary for political decision-makers, commanders, and users to be reasonably held responsible for the results of decisions to deploy or not deploy a system.

However, training alone will be insufficient. How humans use the systems will also be influenced by the user interface and their own psychology. In many instances, the concerns outlined below will need to be treated as ongoing activities that must be considered at most or all of the lifecycle stages. Efforts to address many of the concerns raised at later stages will need to be started at this stage.

Understanding the technology:

1) Policymakers, commanders, and users need more than a basic understanding of the technology. Users will likely need more advanced education—including fields such as computer science, data science, and AI—while military leaders and decision-makers should be trained in AIS technologies. This will involve identifying how many people need to be part of a team overseeing a system, to employ a broad enough technical and programming skill set. The type of training and skill set necessary needs to be determined, and may depend on individual systems.

2) Users and commanders need to understand not only how the technology works, but also how the system interacts with the real world and what will happen if the environment changes. They need to understand the limitations of the system's capabilities, especially those with the most legal and ethical concerns, such as target profiles. Commanders and users need to have a clear understanding of what the system realistically can and cannot do.

3) Procurement officers and people in TEVV are among those through the chain of command and across the lifecycle who will need to have some level of training. Only with in-depth knowledge and understanding of the systems and technology can they be reasonably held responsible for their stage.

4) Militaries need to be prepared to bring on people with the specialized skills necessary to use and understand these technologies.

Understanding the AIS and its "decision-making" capabilities:

1) Users and commanders need to understand what the system was designed for and that they are using it accordingly. This is both a training and design issue. It includes understanding and mitigating the inherent uncertainty associated with AI in systems, and reassessing the use of the system if commanders and/or users are not able to properly account for the uncertainty. Users and commanders will need to be trained to assess the risks and tradeoffs of using the system, especially on legal and ethical issues. What it means for commanders and users to be "well informed" about a system will need to be defined, documented, and assured throughout the lifecycle.

2) Humans can more easily predict the types of mistakes that other humans will make than the types of mistakes that machines will make. Is it possible to train users to anticipate the types of mistakes that

will likely be made by a system they are working with? If not, a determination needs to be made if the AIS considered for use in a military setting is too unpredictable, potentially putting it in violation of IHL.

3) Can tests be developed to familiarize users with the system's capabilities and limitations? Can training be provided to teach users what to look for to determine if the system is functioning erratically or incorrectly in some way? This will be necessary to address well-identified concerns regarding automation bias.

4) Several issues will require a mix of system design, user interface, and human training. This includes enabling commanders and users to understand the speed with which the system is "acting" versus their own human response times; why the system functioned in the ways it did; the causal chain that led to the action; the limitations of the system's ability to "explain" its actions; what will trigger the system to take an action; the scenarios and situations represented in the datasets on which the system was trained, as compared to those in which it might be used; when the system is clearly in the user's control, and when it can no longer be called back; and the responsibility that goes with deploying a system that is beyond their control in different situations and scenarios.

5) Some type of certification will likely need to be developed to certify that users of a system are sufficiently trained to work with its interface and address the issues identified above.

6) Traditional target acquisition processes were designed for humans. Have these been updated—or do they need to be—to reflect differences in how AIS may identify targets? Personnel in tactical units will need to understand how targets are nominated and approved, and how weapons employment is authorized. This includes detailing how a human would provide final approval of weapon employment to avoid violation of IHL.

General questions pertaining to education and training:

1) Should training cover the dual-use nature of system components? If components were not originally designed for the system, how much do users need to understand about the history of the components, how they were originally intended to be used, and what that implies for system weaknesses?

2) An AIS will likely have many components that need to be updated at different times. How much of an understanding of each component do users need, versus an understanding of the system as a whole, to be reasonably held reliable for overseeing and using the system?

3) Should training cover how the user can decline to deploy the system if they are concerned that it will not function as expected? Concerns could be due to problems with the system, or the mission being outside of system parameters. Are there alternative systems the user can deploy, especially in self-defense scenarios where the user may feel they have no other choice?

4) Is there a backup plan if the system cannot be used or if it fails? Is it possible for users to take over the mission and, if so, how?

5) How can the commander and/or user provide assurance that they understood the capabilities of the system and used it correctly, even when the system functioned as intended? During review stages, how can it be confirmed that people were properly trained and used the system correctly, as opposed to acting incorrectly but getting lucky that the system nonetheless happened to function as intended?

6) What training is necessary for commanders and users to understand the physical parameters of the system and not to try to deploy it outside of those parameters?

7) Do users understand the potential ramifications of being the first to use a system, if it has not been used before or elsewhere, or if it has been recently updated?

8) If problems with the system arise during training, is a process in place to send the system back for improvements or updates, or to decide not to use the system?

9) Is there training or knowledge sharing for political leaders and commanders to provide a sufficient understanding of the system's capabilities and limitations to authorize its use?

10) Have the system's users been working with its developers to create human- or soldier-centered design? Do the users understand that, although developers will do their best, they will not be able to anticipate and address every possible issue that could arise?

5.6. POLITICAL AND STRATEGIC CONSIDERATIONS

Political leaders and policy officials should understand the primary differences between AIS and "traditional" weapon systems, especially with respect to the human-system interface, which is necessary to verify the system is used legally.

Political leaders and senior military officials need a deep and common understanding of the costs, benefits, and risks of AIS compared to non-autonomous weapon systems. This includes "normal" system operating parameters, limitations, and the potential consequences of system degradation or failures—including the dynamics of conflict. For example, if uncrewed systems have a lower threshold of use than crewed systems, their deployment may make the use of force more likely. Of course, if the system deployment is determined not to be legal, that would have significant implications for leaders who approve it.

Policy decision-makers and senior military officials need to discuss approval processes for AIS deployment and employment in military contexts and understand both normal and emergency command and control processes and procedures. They need to check that military officials and political leaders have the same understanding of rules of engagement, normal system performance, system limitations, risks, acceptable and unacceptable political risks, risk mitigation measures, potential failure modes and mitigation controls, and assignment of overall responsibility and accountability for system failures.

National leaders and senior military officials must take into account the geopolitical environment and level of tension when considering the deployment and employment of a system. They may elect to notify other states of pending deployment—allies and partners, potential adversaries, and/or third parties. They should have backup plans in place for hacked AIS.

Strategic questions to be addressed early on are whether use is appropriate, and what its implications might be. For example, does the development of such systems incentivize states with more limited capacities to employ less reliable or well-tested AIS for military purposes? What norms of use does this establish, and what implications does that have in other areas, such as export control regimes or the safety of the sea lines of communication? What are the implications for the international arms control and non-proliferation regimes? What are the potential consequences of use in other domains? For example, would use against a state with limited or no military AIS capacity prompt retaliation through cyberattacks on critical infrastructure? Most importantly, how can states best address these types of questions?

Issues to consider:

1) Military organizations, in coordination with appropriate policy officials, should comprehensively analyze the risks and benefits of deploying and employing an AIS versus any other weapon system. They should give special consideration to concerns that may arise from AIS. Senior military officials should discuss with policy decision-makers and make available the results of such analyses.

2) Senior military officials need to brief the appropriate policy officials, government officials, and political decision-makers in advance when potential AIS deployment and employment is considered.

3) Senior military officials and policy decision-makers need to agree on approval processes and delegation of decision-making authorities for AIS deployment and employment.

4) Policy decision-makers should have a sufficient understanding of AIS to make meaningful decisions about approving or disapproving their deployment and employment under specific conditions.

5) Legal reviews should verify to senior military officials and policy decision-makers that the use of an AIS will conform with all national and international laws, regulations, directives, and policies, including IHL/LOAC. These reviews should have occurred and been documented at earlier stages, and may also occur currently. They should include obligations under Article 36 of Additional Protocol I to the Geneva Conventions. They should take into account the potential legal ramifications of degraded system performance or system failures, including how deviations from conditions experienced during testing could affect system performance in unexpected ways. This should also cover challenges attached to "learning" systems.

6) Senior military officials should be able to explain to policy decision-makers the different potential operating modes of an AIS—fully autonomous, semi-autonomous, automated, or controlled remotely—along with the planned operating modes for a given operational mission, and the constraints and criteria for switching between modes.

7) Senior military officials and policy decision-makers should agree on what constitutes acceptable and unacceptable risks when deploying AIS. Military officials should provide details of planned risk mitigation measures.

8) Policy decision-makers should understand the nature and implications of potential system failure modes, if not the exact technical causes. In coordination with senior military officials, they should understand who will be held responsible and accountable for significant failures during AIS employment; the channels of communication in the immediate aftermath of such failures; the processes to be used for accountability; and the timing and content of disclosure to the public, if any.

9) Senior military officials should discuss with policy leaders and political decision-makers the rules of engagement for AIS in peacetime, crisis, and conflict situations.

10) Senior military officials need to coordinate with subordinate military units what the command and control process is for remotely updating AIS that are already deployed. This includes establishing criteria for approval levels and delegation of approvals. The process needs to include legal reviews, as applicable, and procedures for notifying higher-level military and policy officials when updates to fielded AIS could substantially change its operational performance.

11) Based on discussions with senior military officials, political leaders should be prepared to respond to the fallout from AIS failures that lead to civilian casualties, collateral damage, non-combatant targeting, etc. This includes understanding potential worst-case outcomes related to AIS failures and requiring senior military officials to provide mitigation measures.

12) Senior military officials and political leaders need to consider the escalation potential of using AIS during crises and conflicts. This includes an assessment of the likelihood that adversary nations may be more willing to target an AIS than a human-crewed system, what the potential political and military costs and benefits of this may be, and what proportional responses to such actions may look like.

13) Senior military officials and political leaders need to understand the chain of command, and command and control procedures, for AIS employment. This includes the ability or inability to recall a system once deployed, or to terminate weapon employment.

14) Senior military officials and policy decision-makers should discuss how they will communicate to the public about potential AIS deployment, employment, and system failures or inadvertent engagements.

5.7. OPERATIONAL-LEVEL COMMAND AND CONTROL

The “operational level” refers to the command-and-control layer between a strategic headquarters—such as an overall command headquarters—and tactical units responsible for executing military operations. In military language, this operational level is often referred to as “theater-level command and control.” It represents the command and control nodes responsible for planning and managing the overall military campaign within a specific geographic and operational area, and for planning and managing the domain-specific components of the overall joint or combined military campaign.

This stage of the lifecycle applies to military operations involving an operational-level command and control headquarters and domain-specific operations centers. It centers on how to carry out operational-level command and control of AIS. This includes the processes by which operational-level commanders advocate for potential deployment and employment of a system, oversee its operations, maintain awareness of its operating areas, and maintain authority, direction, and control over weapons employment.

This stage assumes prior policy and political approvals to deploy and employ an AIS. Operational-level commanders need to understand normal system performance, system limitations, restraints, constraints, risks, risk mitigation measures, and potential failure modes. This includes the potential for IHL/LOAC violations after a system has failed or taken unexpected actions. Commanders need to understand the system’s rules of engagement and special operating instructions, and promulgate these to operational- and tactical-level units. Subordinate commanders need to understand who is held responsible and accountable for system failures that result in unintended consequences. They need to understand the processes and mechanisms by which the system can be recalled or, if applicable, forced into operating in a non-autonomous mode, and the potential for escalation due to unintended or unplanned engagements.

Issues to consider:

1) Operational-level commanders should require a risk-informed, cost-benefit analysis of AIS versus other “traditional” weapon systems. In coordination with commanders from contributing military services, they should be able to clearly articulate their decision to request AIS based on understanding the technology, advantages, and disadvantages of autonomy over other available systems, risks—including to mission and to force—legal considerations, and geopolitical constraints.

2) Operational-level commanders should receive appropriate education and training on a system's operations prior to systems arriving in a conflict environment. This includes normal operating modes, constraints and restraints, and potential failure modes.

3) Operational-level commanders should understand a system's different operating modes and whether all or some options are available once it is deployed—for example, fully autonomous, semi-autonomous, automated, or controlled remotely. Deployment restraints are set by rules of engagement, and operational-level commanders should provide guidance regarding the criteria, constraints, and decision authority for allowing changes to these operating modes.

4) Operational-level commanders should establish processes and criteria for up-chain and down-chain communications, in order to provide common higher- and lower-level awareness of AIS deployment and potential weapons employment.

5) Operational-level commanders must provide rules of engagement that cover the entire geographic and operational area where military operations take place. This includes any special instructions that govern a system's deployment or employment and use of onboard weapons or targeting guidance. The rules of engagement and special instructions must adhere to higher-level guidance previously established by senior military and policy officials.

6) The targeting process should be tailored according to the specificities of a system once it is deployed.

7) Operational-level commanders need to understand the chain of command, and command and control procedures, for AIS employment. This includes the ability or inability to recall a system once deployed autonomously, or to terminate weapon employment.

8) Operational-level commanders should require a formal legal review of any AIS prior to any potential deployment and operational use. This review should include the potential legal ramifications of possible system failure modes. In coordination with legal advisors and subordinate tactical units, theater commanders should assign appropriate levels of responsibility and accountability so that the use of a system does not result in IHL/LOAC violations. This is congruent with existing guidance for the deployment of non-autonomous systems.

9) Updates to software once a system is deployed may interrupt or change its operation. The system's program office, end-user organization, testers, and operational-level commanders should establish

criteria for updates to deployed systems. A risk-based framework will be useful: for example, low-risk routine updates will likely be approved at lower levels than updates that could result in higher risks, such as changes in targeting or weapons employment processes. Criteria for exceptions need to be commonly understood by all stakeholders. Theater and/or operational commanders should consider including updated criteria and processes in the applicable rules of engagement and special instructions. Processes need to include legal reviews, as applicable, along with procedures for notifying higher-level military and policy officials when updates to a fielded system could substantially change its performance.

Issues that should have been considered at earlier stages, but which are flagged here again for redundancy:

1) Operational-level commanders should have a general—if not granular—picture of where AIS is operating in the geographic area where military operations take place, and in what potential modes: autonomous, semi-autonomous, automated, or remote-controlled.

2) Procedures should have already been established for losing communications with a system, or when an adversary or neutral party has taken control of it. They should include minimizing the risks of loss of sensitive equipment or software; immediately notifying superior military and policy officials; and, as approved by policy decision-makers, communicating directly with military representatives of an adversary or neutral party that has intentionally taken or inadvertently recovered a system.

3) Requirements should have been set for subordinate tactical units to provide appropriate mitigation measures that account for system limitations, potential failure modes, unexpected system behavior, lost communications, etc.

5.8. TACTICAL EMPLOYMENT

At this stage, an AIS has been deployed into an operating area and is prepared to employ its onboard weapons. Primary considerations include “normal” operations; adherence to operational- and other higher-level rules of engagement and special instructions; command and control across all possible operating modes; intervention and/or recall procedures; up-chain notification requirements; remote software update procedures; and procedures for emergencies such as lost communications, or if a system is recovered by an adversary or third party. As with all weapon systems, tactical units require a detailed understanding of “normal” system performance, system limitations, restraints and constraints, risks, potential failure modes, risk mitigation measures, and weapon employment rules of engagement.

Feedback from the tactical unit should be included in the detailed record resulting from each deployment. Assuming previous responsibility and accountability was properly addressed at earlier stages, overall responsibility and accountability for AIS operations and weapons employment resides at this unit level. Many issues that arise below should also have been addressed at earlier stages, but they are addressed here again so the operators and commanders can be more familiar with how the system will function in the specific scenario associated with employment of the weapon.

Issues to consider:

1) Prior to system deployment, tactical units should complete a risk-informed, cost-benefit analysis of its use in comparison to other "traditional" weapon systems. Unit commanders should be able to clearly articulate their decision to recommend an AIS based on a detailed understanding of the technology, advantages, and disadvantages of autonomy over other available systems, risks—including to mission and force—test and evaluation results, potential kinetic and non-kinetic threats and adversarial attacks, IHL/LOAC, rules of engagement, and other relevant geopolitical constraints as provided by headquarters.

2) Tactical-level personnel must receive in-depth training on a system's operations prior to deploying it to an operating area. This includes "normal" operating modes and parameters, constraints and restraints, system response to interrupted or lost communications, potential failure modes or unexpected behavior, and existing policy guidance or directives. The training should include academics, modeling and simulation, and/or high-fidelity simulators, experiments, and training exercises. Personnel within tactical units should understand how deviations in the operational environment beyond the conditions experienced during testing could affect the system's performance once deployed operationally.

3) In coordination with the operational-level command and control node, tactical units must provide detailed guidance for a system's chain of command. This includes establishing overall responsibility and accountability for every facet of system operations and weapons employment. Once established, this guidance should not be altered without coordinating with, and receiving an explicit acknowledgment from, subordinate and superior command and control nodes. Tactical-unit personnel should be able to communicate without prejudice any legal, ethical, safety, or suitability concerns to unit commanders before system deployment.

4) Tactical units should have clear avenues for communicating with software developers if they encounter problems with the interface or software.

5) Tactical-level personnel must understand, have been trained in, and be certified in the different operating modes of a system and which options are available once it is deployed. Based on operational-level guidance, tactical-level personnel should establish processes and procedures for permitting and implementing changes to these operating modes once a system is deployed.

6) Based on operational- and other higher-level guidance, tactical-level personnel should establish processes and criteria for up-chain communications, so there is common awareness of system deployment and potential weapons employment.

7) Personnel within tactical units must understand and adhere to IHL/LOAC, operational-level rules of engagement, and special instructions that govern AIS deployment and employment and use of onboard weapons, especially targeting guidance and post-weapon employment notification procedures. Tactical commanders and unit personnel should be briefed on the results of any formal legal reviews of a system prior to deployment and operational use. Reviews should include the potential legal ramifications of possible failure modes.

8) In accordance with operational-level and other higher-level guidance, tactical units should establish detailed procedures for system command and control and weapons employment. This includes maintaining awareness of the system's location, ability, or inability to recall it once deployed or to terminate weapon employment. These procedures should account for operations in environments where communications are limited or intermittent.

9) Tactical units need to establish procedures that account for lost communications with an AIS, or when an adversary or neutral party has taken control of it.

10) Tactical units must establish mitigation measures that account for limitations, potential failure modes, unanticipated system behavior, interrupted or lost communications, etc.

11) Tactical units should understand the potential benefits and risks of updating system software after the system is deployed to the operational level. Tactical-level personnel must adhere to higher-level guidance on the procedures for remote software updates once a system is deployed at the operational level, including the requirement to notify higher levels in advance of a major software update. Criteria for this will be coordinated between the tactical unit and the primary common and control node within a specific geographic area where military operations take place.

12) Tactical units should understand the criteria under which the system must be deactivated remotely, along with the procedures for doing so. This should include steps to prevent an adversary or third party from retrieving sensitive hardware or software onboard.

13) Tactical-level commanders need to understand why and how they could be held responsible and accountable for system failures or IHL/LOAC violations.

14) Tactical-level commanders need to be aware that the use of AIS in an attempt to reduce casualties for one's own troops could escalate the situation, leading to greater risk to troops later on.

5.9. REVIEW, REUSE, AND/OR RETIRE

In this final stage of the lifecycle, an AIS has been used and is now under review to determine if the system should be reused as is, updated or modified for future use, or retired from use.

An after-action review is a technique for improving process and execution by analyzing the intended and actual outcome of an action, and identifying practices to sustain, improve, or initiate. These practices are then used to implement necessary changes at the next iteration of the action. An after-action review is forward-looking. For an AIS, it should evaluate performance against standards; identify strengths and weaknesses; and decide how to improve.

While a unit that employed a system could conduct its own ad hoc review, an established process should be in place which includes updated risk assessments, and identification of improvements needed before potentially using the system in combat again. Previous lifecycle stages may need to be cycled through again, depending on whether software or hardware updates are made, new rounds of TEVV and human training are necessary, etc.

Questions to be considered at this stage include the following:

1) Did the system function as expected? If so, can it be determined why? If not, can it be determined why?

2) Is it clear who was responsible? What documentation exists to identify responsible parties?

3) Is another round of TEVV needed for the system to be reused? Can this be done by members of the military or contractors, or will an independent group also need to perform an assessment?

4) If the system functioned as intended, does it matter if the commanders and users understand why? Is there some way to determine if commanders and users did not have enough training but got lucky with the system functioning as intended anyway?

5) If the system did not function as intended, to what extent was it a result of technical error, environmental error, operator error, leader decision error, a system not fit for that particular mission, etc.?

6) Was the intent clearly articulated and documented prior to deployment? Did the system's functions satisfy the intent of the mission? If not, what went wrong?

7) If the system had to be recalled or the mission had to be aborted, was it safely returned without taking any undesirable actions? If not, what happened and why?

8) If the system was lost to the enemy, is there any way of determining whether secure information remained secure? What is the contingency plan?

9) What is the feedback process for commanders and users post-mission, and to whom is the feedback sent?

10) What policies are in place for reporting on incidents around cyber, safety, archiving of material from each firing session, etc.?

11) Once a system has been used, how will risk assessments be updated?

12) How often are systems reviewed after use, even if the mission was completed successfully? Will this depend on the cost and size of the system, the complexity of the technology, the size of the team overseeing the system, or other things?

13) Was the trigger for engagement accurately identified in advance?

14) If the system is retired, are data and sensitive hardware made safely redundant?

Other issues to consider:

1) Reviews will have to be recurring because autonomous capabilities need to be regularly updated.

2) AIS may have a range of functionalities, each requiring a different approach for updates.

3) Experiences with one system should be considered if they can usefully inform updates that may be needed for other systems.

4) Simulation should be part of the standard preparation and review of a system. The simulation can be run after employment to enhance other machines by using the most recent operational data.

5) The psychological impact of the use of an AIS on the decision-makers and the military personnel should be assessed and appropriate treatment provided if needed.

6. ACRONYMS

AIS	autonomous and intelligent system
AWS	autonomous weapon system
FD	food defender (the uncrewed underwater vehicle used in the scenario)
IHL	international humanitarian law
IHRL	international human rights law
LOAC	law of armed conflict
TEVV	testing, evaluation, verification, and validation
UN	United Nations
UUV	uncrewed underwater vehicle

RAISING THE WORLD'S STANDARDS

3 Park Avenue, New York, NY 10016-5997 USA http://standards.ieee.org

Tel.+1732-981-0060 Fax+1732-562-1571